radware | SHIELDSQUARE

MALICIOUS BOTS ON PROPERTY PORTALS

Avoiding Common Engineering Pitfalls

Introduction

The *Malicious Bots on Property Portals - Avoiding Common Engineering Pitfalls* eBook is the result of analyses conducted on some of the top property portals *(aka, real estate websites)* in the world. The report outlines various hidden issues found in property portals that can be exploited to scrape data.

Everyday, the presence of bots, both good and bad, on various websites is increasing tremendously. If half of the Web traffic is from bots, more than 70% of them are created with malicious intents.

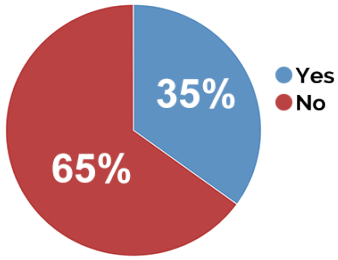
These bad bots can negatively impact websites by degrading SEO rankings, increasing spam, content theft, unwanted network traffic, and even pave way for potential DDoS attacks. These bots are able to scrape by exploiting the vulnerabilities in the website, which may be due to design and technology, or the back-end operations methodology implemented while developing the website.

In this report we have analyzed the most common pitfalls that are found in property portals and how bad bots can exploit these drawbacks to extract data. Most websites have vulnerabilities that will be exploited by bots designed with malicious intent. Just watch out for these common pitfalls outlined in this eBook.

Common Engineering Pitfalls in Property Portals

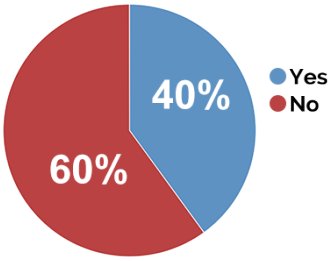
Using Numeric ID or Page Numbers in Incremental Pattern for the Property Listing

Using numeric ID or page numbers in url to access the listing will make it easier for the bad bots to get data from the website. Just by sending requests with page number or ID that is incremented every time, will provide the listing. If the numeric IDs are random then it will be difficult to get valid data. For example, consider that *www.example.com/sale/63896* is a property listing. The ID of the listed property is *63896* and if the IDs are based on increments then the next listing will be having an URL *www.example.com/sale/63897* . The scraper will only require to find the starting ID of the listing and then increment it until he gets what he needs.



35% of the property portals display property ID in the URL

Access to Contact Information of Agent or Seller

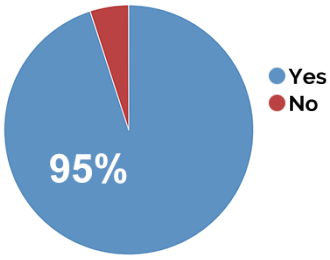


40% of the property portals expose agent/seller contact information

Some property listing websites will also display the contact information of the agent or seller. This information can be misused by other competitors in the property portals business. They will be sending spam mails, newsletters, sign up invitations, offers and unsolicited text messages to the agent or seller. Most customers are very much concerned about the privacy and they do not use certain services that do not ensure privacy. This can result in the customer leaving the current service and subscribing to another that's high on privacy.

URL Following Same Pattern

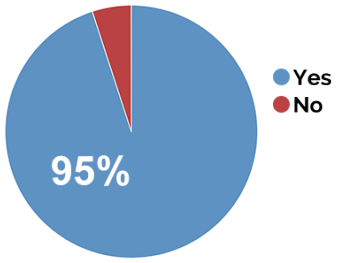
Most of the property listing websites tend to follow a particular pattern for their listing URL. This makes it very easy for the scraper to collect the required data. For example, consider the url www.example.com/forsale/ca/sacramento/95610/5431489_id



95% of the property portals have URLs following the same pattern

In the above url, CA (*California*) is the name of the state, Sacramento is the county, *95610* is the area zip code and *5431489_id* is the property ID. Due to this pattern used in the url, if a scraper wants only the details of all the properties of a particular county, then he can use url format www.example.com/forsale/ca/county_name. In this URL, the scraper will be substituting the county name with that of one which he wants and get the details of properties. An iterative approach is used to get the necessary data. Regular expressions can also be used to generate the rest of the url.

Simple HTML Design with Constant XPATH for all the Pages

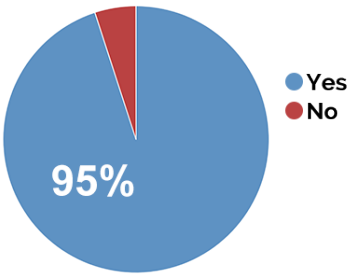


95% of the property portals have HTML with constant XPATH for all pages

Using simple HTML designs will make it easy to extract data in a structured manner and to filter data. If all the pages follow the same design, then the XPATH also will be same. This means that the scraper has to use only one single XPATH, making his job very easy. A typical property portal listing will contain price, property location, area, build date and other features. All these attributes are common and will be repeatedly used in every listing. These attributes are targeted by the scrapers. When these elements are positioned in the same location on the web page, a constant XPATH can be used for getting each of these attributes which makes it vulnerable against bots that extract particular data. Use of dynamic designs for various pages make bulk data scraping difficult.

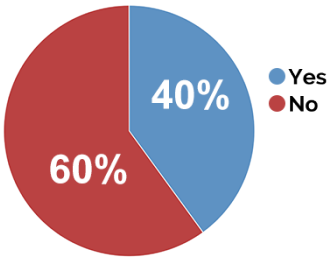
Search Results Provide Direct URL Link to the Listing

Searching for a property is based on various parameters like area, cities, price range, builder, amenities, and so on. When all the parameters for searching are activated, then the website tends to provide all the listings in the search. Many websites are a potential target for scrapers when regular expressions are used in the search box. Most of them tend to list out all the posts as search results. If the search results are providing static URL's to the property page then the scraper only has to scrape the search results to get the URLs of all listings. This can be avoided by using dynamic URLs with redirection in the search results.



95% of the property portals exhibit static URL search pitfall

Usage of Geographical attributes (*ZIP codes, latitude:longitude*) to access listing

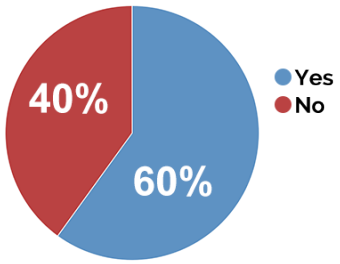


40% of the property portals have the geo-value pitfall

Many websites use geographical attributes like zip codes, latitude and longitude for listing the properties available in that particular location. A user can search using zip codes, area name or pointing the location on map. As you know, the geographical coordinates/ attributes will be constant for a particular area throughout the world. Hence a scraper can easily get all these geographical attributes that are used in the website. They can then apply a brute force method using these values to scrape the listings.

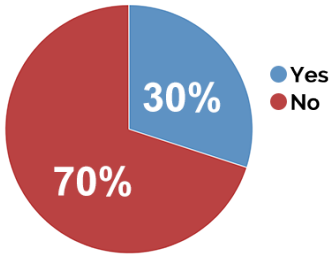
Providing Static URL of the Listing in Sitemap

Sitemaps are used to increase the search engine visibility for the contents of the websites. Search engine spiders use the sitemaps to gather the data from all the pages. Like the search engines even bad bots are able to access the sitemaps. But most of their websites update their sitemaps whenever a new property has been listed. This will make it easier for the scraper to get the data as he will be knowing whenever the updates is made and also he is having direct access to all the URL's of the properties listed in the website.



60% of the property portals have their sitemaps exposed

Inefficient Use of Cookies

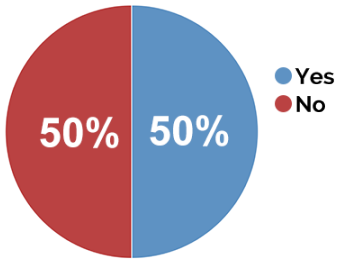


30% of the property portals have inefficient cookie usage

Cookies are one of the best ways to get information about the website visitor. These cookies can contain information like what kind of properties the user was looking for. Most users will be looking for properties from a particular area or within a particular budget. This will help verify if the users are genuine. However, if the searches made are mostly unrelated, it will be an indication of suspicious activities happening on the website. Efficient usage of cookies can deter scrapers to an extent. Passing parameters through cookies and using uniques ID will be a good defensive measure for the website.

Similar Designs for Multiple Websites

After analyzing the security pitfalls of several top property portal websites, it has been understood that most of them follow a similar type of web design. Following similar kind of design makes it easier for the bots to scrape data, since only a single bot is required to be designed for all of the websites, with minor tweaks. It is always good to follow unique designs for each website.



50% of the property portals use similar design principles

What's Next After Addressing the Pitfalls

Running a property portal takes significant resources, planning, execution, and most importantly, the strategy to outclass the competition. When you're focused on these bigger goals, it's easy to miss out on the aforementioned pitfalls. Even if these pitfalls are taken care of by your in-house bot prevention team, and the scrapers are deterred to some extent, you must accept the fact that they can always come back with a different strategy to scrape data off of your website. The creators of these bot programs eventually improvize the algorithms to circumvent the security fixes in place. This significantly increases the in-house team's workload, and in most cases, the team will be playing catch-up with the scrapers and their quickly evolving scraping techniques. Simply put, fixing vulnerabilities and pitfalls doesn't address bot issues thoroughly.

How then can you protect your property portal and uphold your competitive advantage? The answer lies with complementing the efforts of your in-house team with an automated bot prevention solution so that your website is continuously guarded against malicious bots and scrapers.

About Radware

Radware® (NASDAQ: RDWR), a leading provider of cyber security and application delivery solutions, acquired ShieldSquare in March 2019.

Radware® (NASDAQ: RDWR) is a global leader of cybersecurity and application delivery solutions for physical, cloud and software-defined data centers. Its award-winning solutions portfolio secures the digital experience by providing infrastructure, application and corporate IT protection and availability services to enterprises globally. Radware's solutions empower more than 12,500 enterprise and carrier customers worldwide to adapt quickly to market challenges, maintain business continuity and achieve maximum productivity while keeping costs down. For more information, please visit www.radware.com

Radware encourages you to join our community and follow us on: Radware Blog, LinkedIn, Facebook, Twitter, SlideShare, YouTube, Radware Connect app for iPhone® and our security center DDoSWarriors.com that provides a comprehensive analysis of DDoS attack tools, trends and threats.

Our Data Centers Around the World

Amsterdam,NL
Chennai,IN
Dallas,TX,US
Frankfurt,GR
Hong Kong
Houston,TX,US
London,UK
Melbourne,AUS
Milan,IT
Montreal,CA
Mumbai,IN
Paris,FR
Querétaro,MX
San Jose,US
Sao Paulo,BR
Seattle,WA,US
Singapore
Sydney,AUS
Tokyo,JP
Toronto,CN
Washington D.C,US